DOCUMENT IMAGING AND DOCUMENT MANAGEMENT
PRE-CONFERENCE WORKSHOP
Saturday, September 30, 2000 and Sunday, October 1, 2000

Steve Gilheany
Archive Builders
1209 Manhattan Ave.
Manhattan Beach, CA 90266
stevegilheany@archivebuilders.com

Please note: The workshop notes and slides are available at [http://www.ArchiveBuilders.com]. The technology of document imaging and document management is moving from being an IS (Information Systems) and RIM (Records and Information Management) tool to digital libraries. The technology of scanning, importing, transmitting, organizing, indexing, storing, protecting, locating, retrieving, viewing, printing, and preserving documents requires some study and understanding before a librarian can say with some certainty that a proposed application of the technology will meet the integrity and service requirements for a given library. There are many technological issues, of importance to libraries, just below the surface of document imaging and document management. For example, the best copy of a book is the typesetter file that the book was created from. Another example is the mathematics behind graphic and text of computer-generated documents such as this document: the Proceedings of the IAMSLIC 2000 Convention, Tides of Technology (International Association of Aquatic and Marine Science Libraries and Information Centers). The equations used to define the graphics and text in almost all computer-generated documents is the same mathematics that now defines maps, computer animations, special effects in motion pictures, computer models, textiles, machines, buildings, and models of molecular structures. Future advances in technology will likely be so large that it will appear that there were no changes in the past, just as the invention of pagination in the 15th century does not seem to be a significant change from today's perspective, in the 21st century. The coming of nanotechnology, in thirty to forty years, promises to sweep aside all existing technology, leaving only the record of our thoughts and actions (the content of libraries) as the only significant thing carried from the past, for use in the future [http://www.Foresight.org]. The purpose of libraries is unchanged by the technology of document imaging and document management, which merely provides new tools in support in the traditional purpose of libraries and the traditional purposes of allied institutions such as archives. The following are a few specific examples of document management and document imaging metrics and technology:

Computer storage requirements for various digitized document types (Estimates are rounded and adjusted for ease of use.)

Scanned Letter Size Pages (All images are scanned 1 bit per pixel, black & white, and compressed, unless otherwise noted.)

1 scanned page (8 1/2 by 11 inches, A4) = 50 KiloBytes (KByte) (on average, black & white, CCITT G4 compressed)

1 file cabinet (4 drawer) (10,000 pages on average) = 500 MegaBytes (MByte) = 1 CD (Compact Disc) (ROM or WORM)

2 file cabinets = 10 cubic feet (cf) = 1,000 MegaBytes = 1 GigaByte (GByte); 10 file cabinets = 1 DVD-R (WORM) (see below)

2,000 file cabinets = 1,000 GigaBytes = 1 TeraByte (TByte); 2,000 file cabinets = 200 DVDs

1 box (in inches: 15 1/2 long x 12 wide x 10 deep) (400 x 300 x 250 mm) (2,500 pages) = 1 file drawer = 125 MegaBytes

1 box (packed) = 2 linear feet (500 mm) of files (loose enough for active filing) = 25 (rounded) linear inches = 125 MegaBytes

1 linear inch (~20 mm) = 100 pages = 5 MegaBytes; 1 thousand linear inches = 100 thousand pages = 5 GigaBytes

1 cubic foot (cf) (~.025 cubic meter) = 2000 pages = 100 MegaBytes; 10 cubic feet (~.25 cubic meters) = 20 thousand pages = 1 GigaBytes

8 boxes = 16 linear feet = 2 file cabinets = 1 GigaByte; 8,000 boxes = 16,000 linear feet = 1,000 GigaBytes = 1 TeraByte

1 library book (average, scanned in black and white) = 10 MegaBytes; 50 books = 500 MBytes = 1 CD; 100 books = 1 GByte

For paper and microform document imaging, see also AIIM (Association for Information and Image Management) [http://www.AIIM.org]

For records and information management, see also ARMA (Association of Records Managers and Administrators) [http://www.ARMA.org]

Scanned Microforms

1 roll of 16 mm microfilm (100 ft, ~30 meters) (24X reduction) = 2,500 letter size images = 1 box = 1 file cabinet drawer = 125 MegaBytes

1 roll of 35 mm microfilm (100 ft) (12X reduction, open spacing, normal scan) = 1,000 letter size images = 50 MegaBytes

1 microfiche (105 mm film) (24X reduction) = 100 letter size images = 5 MegaBytes (average); 200 microfiche = 20,000 images = 1 GigaByte

In many record series, each microfiche contains only a few images because each fiche represents a single record in the series (e.g. one fiche per person in a personnel record series). In this case filming breaks on records, rather than being continuous. To a lesser extent this is also true for roll film. In these cases, the amount of storage required depends on the number of images on the film, not the number of microfiche or the number of rolls of film. A full, standard 24X microfiche has 7 rows of 14 letter size (8 1/2 x 11 or A4) images for a total of 98 images.

As with any microform, scanned aperture card images require the same storage as images scanned from the paper original of the document in the aperture.

Compression: All documents are stored and transmitted in compressed format. All compression formats are assumed to be lossless or used with a lossless setting, except MPEG (Moving Picture Experts Group), unless otherwise stated. Lossless or non-destructive compression (as opposed to lossy or destructive compression) does not change the document. That is, a decompressed document is identical to the original document before compression was done. Lossless compression is often needed to meet legal requirements for document storage. The most common form of one bit (per pixel), bitonal (The two tones of color are two shades of gray which are black and white.), lossless compression, used in TIFF G4 and Adobe PDF (Portable Document Format), is the CCITT G4 (Group 4) facsimile compression format. Before using any other form of compression, it is often useful to evaluate the cost savings of moving to the less common format. The CCITT (Comité Consultatif International pour le Télégraphe et le Téléphone) (International Telegraph and Telephone Consultative Committee) is now a part of the ITU (International Telecommunications Union) http://www.ITU.int]. The G4 ITU recommendation T.6 (11/88), Facsimile coding schemes and coding control functions for Group 4 facsimile apparatus, is on pages 48-57 of the CCITT Blue Book, Volume VII - Fascicle VII.3, Terminal Equipment and Protocols for Telematic Services, Recommendations T.0 - T.63, ISBN 92-61-03611-2

Scanned Engineering Drawings / Large Format Documents

1 E size drawing (48 inches by 36 inches) (A0 size) = 16 letter size pages (8 1/2 by 11 inches, metric A4) = 800 KiloBytes. To place an E size drawing in a file folder in a file cabinet drawer, the drawing must be folded in half 4 times and is then 16 sheets of paper thick when folded.

NB: Scanning must accommodate the older, untrimmed, US paper sizes, because it is the older drawings that are digitized by scanning.

Metric Trimmed Paper Sizes					United States Paper Sizes				Storage
Metric Name	**Metric Size in Milli-meters**	**Size in Inches**	**Number of Square Meters**	**Number of A4 Size Pages**	**US Name**	**New Size (Trimmed) in Inches**	**Old Size (Un-trimmed) in Inches**	**Equiv-alent Letter Size Pages**	**Digital Image Storage Requirements**
A8	52 x 74	2.07 x 2.91	1 / 256	1 / 16	Business Card				5 KiloBytes
A7	74 x 105	2.91 x 4.13	1 / 128	1 / 8	3 x 5	3 x 5			10 KiloBytes
A6	105 x 148	4.13 x 5.83	1 / 64	1 / 4	Microfich e				
A5	148 x 210	5.83 x 8.27	1 / 32	1 / 2	5 x 8	5 x 8			25 KiloBytes
A4	210 x 297	8.27 x 11.69	1 / 16	1	A	8 1 / 2 x 11	9 x 12	1	50 KiloBytes
A3	297 x 420	11.69 x 16.54	1 / 8	2	B	11 x 17	12 x 18	2	100 KiloBytes
A2	420 x 594	16.54 x 23.39	1 / 4	4	C	17 x 22	18 x 24	4	200 KiloBytes
A1	594 x 841	23.39 x 33.11	1 / 2	8	D	22 x 34	24 x 36	8	400 KiloBytes
A0	841 x 1189	33.11 x 46.81	1	16	E	34 x 44	36 x 48	16	800 KiloBytes
2A0	1189 x 1682	46.81 x 66.22	2	32					1.6 MegaBytes
					US Roll Sizes				
					F	28 x 40	varies		600 KiloBytes
					G	11 x (22 1 / 2 to 90)	varies		
					H	28 x (44 to 143)	varies		
					J	34 x (55 to 176)	varies	Sizes G, H, J, and K	
					K	40 x (55 to 143)	varies	are US roll sizes	

Paper size references: MIL-M9868-D, Microfilming of Engineering Documents, 35MM, Requirements for, 10-1-70 and amendments 1 and 2, 2-12-82 and 9-20-82; MIL-STD-804B Format and Coding of Aperture, Copy and Tabulating Cards Engineering Data Micro-reproduction System, 15 August, 1966; ANSI Y 14.1, 1980, Drawing Sheet Size and Format, published by ASME (American Society of Manufacturing Engineers), New York; Metric standards first published in 1922 by DIN (Deutsches Institut für Normung) (German Institute for Standards) [http://www.DIN.de] Now used worldwide as ISO 216. [See last page for ISO reference.]

Newspapers: A double truck (center fold) full broadsheet is 24 in x 36 in, equivalent to an old D size drawing in size. Because a newspaper page would be scanned at a higher resolution and contains detailed graphics, a double truck would require about 1 MegaByte and a single full broadsheet page (18 by 24 inches) would require about 1 / 2 MegaByte. See also NAA (Newspaper Association of America) [http://www.NAA.org]

DVD Digital Video Disc

See also [http://www.DVDdemystified.com/dvdfaq.html]

DVD (commonly Digital Video Disc) (A DVD is the same physical size as a CD.) DVD stands for Digital Versatile Disc, by vote of the committee that controls the trademark DVD, the DVD Forum. [http://www.DVDForum.org] All capacities are given in commercial units: e.g.: 1 GigaByte = 1 Billion Bytes; 1 MegaByte = 1 Million Bytes

NB: When you calculate the amount of storage you will need on a given CD or DVD (using the table below), be sure that the units you are using for the size (amount) of data you plan to record are given in commercial rather than computer units. If you are not sure that the size (amount) of your data is given in commercial units, then add 10 (ten) percent to the size (amount) of data you plan to record. In all cases, you should leave yourself some headroom (of at least 5 percent) for last minute changes. (This can be reduced as you gain experience.) If, in addition to the normal headroom allowance, you are also uncertain of the (data size) units used, it is best to allow a total of 15 percent for headroom.

Metric Trimmed Paper Sizes					United States Paper Sizes				Storage
Metric Name	**Metric Size in Milli-meters**	**Size in Inches**	**Number of Square Meters**	**Number of A4 Size Pages**	**US Name**	**New Size (Trimmed) in Inches**	**Old Size (Un-trimmed) in Inches**	**Equiv-alent Letter Size Pages**	**Digital Image Storage Requirements**
A8	52 x 74	2.07 x 2.91	1 / 256	1 / 16	Business Card				5 KiloBytes
A7	74 x 105	2.91 x 4.13	1 / 128	1 / 8	3 x 5	3 x 5			10 KiloBytes
A6	105 x 148	4.13 x 5.83	1 / 64	1 / 4	Microfich e				
A5	148 x 210	5.83 x 8.27	1 / 32	1 / 2	5 x 8	5 x 8			25 KiloBytes
A4	210 x 297	8.27 x 11.69	1 / 16	1	A	8 1 / 2 x 11	9 x 12	1	50 KiloBytes
A3	297 x 420	11.69 x 16.54	1 / 8	2	B	11 x 17	12 x 18	2	100 KiloBytes
A2	420 x 594	16.54 x 23.39	1 / 4	4	C	17 x 22	18 x 24	4	200 KiloBytes
A1	594 x 841	23.39 x 33.11	1 / 2	8	D	22 x 34	24 x 36	8	400 KiloBytes
A0	841 x 1189	33.11 x 46.81	1	16	E	34 x 44	36 x 48	16	800 KiloBytes
2A0	1189 x 1682	46.81 x 66.22	2	32					1.6 MegaBytes
					US Roll Sizes				
					F	28 x 40	varies		600 KiloBytes
					G	11 x (22 1 / 2 to 90)	varies		
					H	28 x (44 to 143)	varies		
					J	34 x (55 to 176)	varies	Sizes G, H, J, and K	
					K	40 x (55 to 143)	varies	are US roll sizes	

Paper size references: MIL-M9868-D, Microfilming of Engineering Documents, 35MM, Requirements for, 10-1-70 and amendments 1 and 2, 2-12-82 and 9-20-82; MIL-STD-804B Format and Coding of Aperture, Copy and Tabulating Cards Engineering Data Micro-reproduction System, 15 August, 1966; ANSI Y 14.1, 1980, Drawing Sheet Size and Format, published by ASME (American Society of Manufacturing Engineers), New York; Metric standards first published in 1922 by DIN (Deutsches Institut für Normung) (German Institute for Standards) [http://www.DIN.de] Now used worldwide as ISO 216. [See last page for ISO reference.]

SS (Single Sided), DS (Double Sided), SL (Single Layer), DL (Double Layer), SS/SL (Single Sided / Single Layer), DS/DL (Double Sided / Double Layer), DS/SL (Double Sided / Single Layer per side), DS/ML (Double Sided / Mixed Layer; one side 1 layer, other side 2 layer), HD (High Density); Top Layer (Layer 1), Bottom Layer (Layer 0)

* CD capacities have always been advertised as 650 MegaBytes using the older computer based MegaByte (1,048,576 Bytes) size. Using the new commercial standard units of 1 Million Bytes per MegaByte, a CD holds 682 MegaBytes. DVD capacities, however, are always stated in the new, smaller, commercial units.

** DVD-R (Recordable) and CD-R are the equivalent of WORM (Write Once, Read Many) The DVD-R capacity listed above, of 4.7 GigaBytes per side, is for discs and DVD writers that conform to the new DVD-R standard (DVD-R 2.0). The new DVD-R discs and writers are available now. The older (DVD-R 1.0) capacity of DVD-R discs is 3.95 GigaBytes per side for a total of 7.9 GigaBytes for a two sided disc. 25212v005

The workshop notes and slides are available at [http://www.ArchiveBuilders.com].